EDGEWOOD CHEMICAL BIOLOGICAL CENTER
U.S. ARMY RESEARCH, DEVELOPMENT AND ENGINEERING COMMAND
Aberdeen Proving Ground, MD 21010-5424

ECBC-TR-1259

A COMPARISON OF QSAR BASED THERMO AND WATER SOLVATION PROPERTY PREDICTION TOOLS AND EXPERIMENTAL DATA FOR SELECTED TRADITIONAL CHEMICAL WARFARE AGENTS AND SIMULANTS

Jerry B. Cabalo

RESEARCH AND TECHNOLOGY DIRECTORATE

Craig K. Knox

LEIDOS
Gunpowder, MD 21010-0068

July 2014

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorizing documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-07-2014	Final	Jan 2014 - Apr 2014

4. TITLE AND SUBTITLE
A Comparison of QSAR Based Thermo and Water Solvation Property Prediction Tools and Experimental Data for Selected Traditional Chemical Warfare Agents and Simulants

5a. CONTRACT NUMBER
Internal ECBC Funding

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Cabalo, Jerry B. (ECBC); and Knox, Craig K. (Leidos)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Director, ECBC, ATTN: RDCB-DRI-I, APG, MD 21010-5424
Leidos, P.O. Box 68, Gunpowder, MD 21010-0068

8. PERFORMING ORGANIZATION REPORT NUMBER
ECBC-TR-1259

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT-LIMIT 200 WORDS
There is a need for rapid prediction of the physico-chemical properties of chemical warfare agents (CWAs) and toxic industrial chemicals (TICs) on environmentally relevant materials, personal protective equipment, and human tissue. It is the objective of this report to survey the reliability from several available in-silico tools for a set of physico-chemical properties that impact prediction of environmental fate of a set of traditional CWAs and simulants. The tools included EPI Suite, ACD Labs, ADF COSMO-RS, ChemAxon's Marvin, and Vega. Of the predictive tools surveyed, EPI Suite and ACD Labs consistently had the highest accuracy for boiling point, vapor pressure. EPI Suite and ACD Labs gave reasonable results for octanol-water partitioning coefficient and water solubility for most compounds evaluated. For available measurements, ACD Labs, COSMO-RS, and Marvin were within two log units of the measured value.. Arrangements of atoms outside the model training set accounted for much of the error between prediction and experiment. For molecular properties dependent on descriptors such as dipole moment, the effect of molecular symmetry most likely accounts for significant overestimation. The fragment based models are purely additive and molecular symmetry may cause the effect of a pair of functional groups to cancel rather than add.

15. SUBJECT TERMS

pKow	Solubility	Boiling point	pKa
Vapor pressure	Property estimation	Traditional agents	Simulants

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Renu B. Rastogi
U	U	U	UU	36	19b. TELEPHONE NUMBER *(include area code)* (410) 436-7545

Blank

EXECUTIVE SUMMARY

There is a need for rapid prediction of the physico-chemical properties of chemical warfare agents (CWAs) and toxic industrial chemicals (TICs) on environmentally relevant materials, personal protective equipment, and human tissue. While in the past it was possible to concentrate laboratory characterization efforts on a limited number of known, traditional CWAs and TICs, there is the possibility that state and non-state actors may use CWAs outside of the traditional CWAs that have distinctly different physical and chemical properties. Rapid, reliable hazard assessments for the persistence and spread of non-traditional agents may be necessary for the benefit of first responders and clean up teams before laboratory measurements can be done. Predictive tools also serve as screening tools that help identify compounds that may be particularly difficult to decontaminate. It is the objective of this report to survey the reliability and error spread from several available in-silico tools for a set of physico-chemical properties that impact prediction of environmental fate of a set of traditional CWAs and simulants. Except for ADF COSMO-RS, the tools are based on quantitative structure-activity/property relationship (QSAR/QSPR) methods using molecular fragments (group-contribution) approaches, which make predictions based on a regression of laboratory measurements performed on similar chemicals and the underlying statistical correlations that describe the property variations resulting from specific groups of atoms within each compound.

EPI Suite, ACD Labs, Marvin, Vega, and COSMO-RS were used to predict properties such as the boiling point, vapor pressure, log of the water/octanol partitioning coefficient (K_{ow}), water solubility, and pK_a. For boiling point, both ACD Labs and EPI Suite were accurate to within 20^o C and 29^o C. EPI Suite, ACD Labs, and ADF COSMO-RS performed quite well for vapor pressure predictions, except that ACD Labs could not generate predictions for vapor pressures of less than 0.1 Pa. For K_{ow}, EPI Suite, ACD Labs, ChemAxon's Marvin, Vega, and ADF COSMO-RS were evaluated, and except for COSMO-RS, the difference between estimation values and measurement were less than one log unit. With respect to water solubility, EPI Suite's K_{ow} estimation method of solubility yielded the smallest average difference of 0.87 log units between experiment and measurement, although ACD Labs could give predictions over a range of temperatures and pH. The greatest difference between prediction and experimental measurement occurred for ADF COSMO-RS presumably because, although it is partially based on accurate Density Functional Theory (DFT) calculations, its overall prediction employs an empirical fit based on an insufficiently small training set size of only 642 compounds. For pKa estimations, experimental data for only three of the compounds examined was readily available in the literature, and so the accuracy of ACD Labs and ChemAxon's Marvin towards the traditional agents could not be adequately evaluated.

Based on the brief survey of estimation methods, both EPI Suite and ACD Labs gave excellent results for boiling point and vapor pressure. For K_{ow} estimations, EPI Suite, ACD Labs, Marvin, and Vega gave estimations that were reasonable and on average less than one log unit off the published measurement. Larger errors were encountered with water solubility and pKa estimations. We reason that two issues contribute to the difference between experimental measurements and estimations from fragment based methods. First, as would be expected, compounds containing elements or functional groups outside of the method's training set

contributed to the average error. The second issue that seems apparent in examination of the data is that molecules that tend to produce the largest differences between model prediction and experimental measurement have molecular symmetry. For properties highly dependent on molecular structure and polarity, such as water solubility, a fragment based method can contribute significantly to the error, since the fragment contributions are treated additively. It is possible symmetry may cancel out the contributions from a given fragment, so that a property is overestimated. We recommend caution with respect to estimations from fragment based methods for molecules that possess symmetry, or possess unusual functional groups or atomic linkages (e.g. N-P bond). We expect that estimation methods based on descriptors of the entire molecule rather than fragments should be more robust with respect to symmetry and functional groups outside of the method training set of data.

Examples of Property Predictions That Differ the Most From Experiment

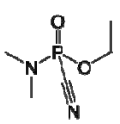
Water Solubility (mg/L)				
	EPI Kow	EPI Frag	ACD	EXP
HN1 (nitrogen mustard)	*4.0E+04*	*7.3E+03*	*1.5E+04*	1.6E+02
diisopropyl methyl phosphonate	7.3E+03	*2.2E+05*	*3.4E+04*	1.5E+03
GA	3.2E+04	*1.0E+06*	3.5E+05	9.8E+04
GB	*4.6E+04*	1.0E+06	4.2E+05	1.0E+06
GD	*1.6E+03*	*3.4E+05*	5.0E+04	2.1E+04
GF	2.1E+03	*5.4E+05*	6.2E+04	3.7E+03
L1	2.6E+02	*4.7E+03*	n/a	5.0E+02
VX	*3.2E+03*	9.1E+04	4.8E+03	3.0E+04
disperse red-9	6.8E-01	*8.6E+00*	4.5E-04	1.2E-01

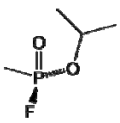
pKow					
	EPI	ACD	MARVIN	Vega	EXP
HD (sulfur)	*2.4*	2.1	2.0	3.2	1.37
VG	1.7	*2.9*	1.8	1.7	1.7
disperse red-9	4.1	*3.0*	*3.0*	*3.0*	4.1

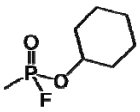
	pKa			
	ACD	ADF/COSMO	Marvin	EXP
VX	*9.8*	7.9	*10.6*	8.6

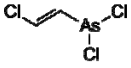
Illustrative Tables: Compounds and predicted measurements highlighted in italics show predicted values of water solubility (mg/L), pK_{ow} and pKa that differ from experimental measurements by more than an order of magnitude or 1.0 pK units. These molecules (some shown below) tend to have a high degree of symmetry or unusual elements or combinations of atoms.

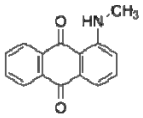

HN1

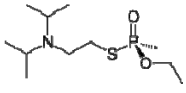

GA


GB


GF


L1


disperse red-9


VX

PREFACE

The work described in this report was authorized under internal U.S. Army Edgewood Chemical Biological Center funding. This work was started in January 2014 and completed in April 2014.

The use of either trade or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

The text of this report is published as received and was not edited by the Technical Releases Office, U.S. Army Edgewood Chemical Biological Center.

This report has been approved for public release.

Acknowledgments

The authors would like to acknowledge Jeffrey McGuire for lending the license for the ACD Labs Software. We also acknowledge helpful comments from Bruce King and Morgan Minyard about the weaknesses and insufficiencies of existing predictive tools towards emerging chemical threats.

Blank

CONTENTS

TABLES

A COMPARISON OF QSAR BASED THERMO AND WATER SOLVATION PROPERTY PREDICTION TOOLS AND EXPERIMENTAL DATA FOR SELECTED TRADITIONAL CHEMICAL WARFARE AGENTS AND SIMULANTS

1. INTRODUCTION

1.1 The Need for a Rapid Prediction Capability

There is a need for rapid prediction of the physico-chemical properties of chemical warfare agents (CWAs) and toxic industrial chemicals (TICs) on environmentally relevant materials, personal protective equipment, and human tissue. While in the past it was possible to concentrate laboratory characterization efforts on a limited number of known, traditional CWAs and TICs, there is the possibility that state and non-state actors may use CWAs outside of the traditional CWAs that have distinctly different physical and chemical properties. Due to the toxicity of many potential compounds not within the list of traditional agents, and because of the difference in behavior on different environmental media, it is not feasible to perform laboratory measurements of all compounds of interest on all possible environmental media.[1] A similar issue confronts government regulatory agencies such as the Environmental Protection Agency (EPA),[2] where the vast number of compounds produced by industry exceeds their laboratory capacity to characterize every possible compound. However, predictive tools help prioritize[3] compounds of interest and target compounds that may have properties that contribute to persistence in the environment or properties that impede decontamination. First responders and clean-up teams may require rapid, reliable estimations of contamination area and penetration into materials before any laboratory property measurements can be done, as well.

The CWA physico-chemical properties contribute to complex but critical processes such as environmental fate,[4-10] pathways into the body,[5, 11-12] as well as the innate toxicity of the compound, and all of these factors contribute to the overall threat. Some examples of physico-chemical properties of importance[13] include solubility in water,[14-16] ionizability in water (pK_a),[17-20] vapor pressure,[21-23] boiling point, and partitioning coefficients,[5] such as K_{ow}, the partitioning coefficient between octanol and water. Solubility in water can affect environmental transport of a given CWA, and whether the compound can undergo degradation by hydrolysis. Ionizability in water is related, and also contributes to whether a compound can degrade in the environment. Vapor pressure and boiling point affect persistence. Sarin has a relatively high vapor pressure and lower boiling point, and tends to evaporate from an affected area within hours. In contrast, VX and sulfur mustard have low vapor pressures and can persist for long periods of time. In addition, low vapor pressure increases the difficulty of detection. K_{ow} has been known to be an indicator for the pathway into the body, where a high value shows a cutaneous threat. Such properties are also indicators of whether the compound penetrates personal protective equipment. Furthur complicating matters, the interaction of CWAs and common materials in the environment such as concrete, sand, and soil, affect the fate of the CWAs and whether the threat from that agent persists over time. As a result, it is not feasible to

experimentally measure all possible combinations of threatening materials and environmental substrates, and there is an acute need for predictive in-silico tools. In other words, reasonably accurate in-silicopredictive tools can augment cost limited laboratory resources by identifying the compounds most likely to have threatening properties.

1.2 Inputs for Tools for Predicting Environmental Fate

Because of the need for in-silico tools that predict environmental fate of toxic compounds, the focus of this study is to examine a number of existing tools that predict physico-chemical properties. The property prediction in turn can be used as inputs into larger scale models that predict environmental fate. There are a number of methods and models for predicting environmental fate of toxic chemicals, particularly pesticides. Because of the similarity of many CWAs to common pesticides, it is possible to leverage these tools. Modeling tools such as PEARL[24] and HYDRUS,[25] make predictions of environmental transport and degradation using the physico-chemical properties of a target compound as inputs. These tools predict the persistence of a compound in the environment, whether it can contaminate ground water, or its behavior in various types of soil, etc. Such predictions are essential for assessing the long term threat, but these tools rely on accurate predictions of physico-chemical properties to obtain reliable environmental fate predictions.

A number of Quantitative Structure-Activity Relationship (QSAR) based software tools currently exist, such as EPI Suite, VEGA (a component of CAESAR), ACD Labs Suite, ChemAxon MARVIN, and SPARC, that utilize geometric/functional group-contribution descriptors to predict physical properties that feed into the fate models. Some are freely available, such as EPI Suite, while others are commercially licensed. Given the variety of predictive tools, an assessment of how these tools perform compared to experimental data is desirable.

.

1.3 Objective

It is the objective of this report to survey the reliability and error spread from several available in-silico tools for a set of physico-chemical properties that impact prediction of environmental fate. We specifically hope to examine performance against a set of traditional CWAs and simulants. Because the environmental fates of other organic compounds are also relevant, such as industrial dyes and pharmaceuticals, we include Disperse Red 9 and cocaine. Some of the tools, as mentioned above are based on fragment based, that is, group-contribution QSAR approach, which makes predictions based on a regression of laboratory measurements performed on similar chemicals. This QSAR approach has been well established for predicting general physico-chemical properties,[26-30] especially in the pharmaceutical industry.[31] A method exists that makes predictions from a descriptor calculated from density functional theory (DFT), which is an electronic structure method, and we include results from the COnductor-like Screening MOdel for Real Solvents[32-33] (COSMO-RS) for comparison. We do not intend to fully analyze the results of COSMO-RS since such a detailed examination of the electron density around the molecule is outside the scope of this study. Although typical performance studies involve

hundreds or thousands of compounds, we wish to limit the scope to a set of traditional CWAs and simulants, and perhaps provide a guide for usage of existing predictive tools in the study of compounds related to traditional CWAs.

2. BACKGROUND

2.1 Overview of QSAR Methods Used

Four of the physico-chemical predicting tools examined in this report are QSAR based tools and a DFT electronic structure based tool. QSARs are a well established method of property prediction first demonstrated for petroleum components. QSARs are simple mathematical regression models of the form

$$Y_{pred} = c_0 + c_1X_1 + c_2X_2 + \cdots + c_{1,2}X_1^mX_2^n + \cdots \qquad \text{Equation (1)}$$

where Y_{pred} is the predicted property, c_0 is a constant, c_1 to c_n are coefficients from the regression to the training set of measurements, X_1 to X_n represent molecular or fragment or field-based descriptors, and the final term in Equation 1 represent higher order terms. The descriptors are some property or characteristic of the molecular structure. Table 1 shows some of the classes of descriptors as well as some examples of those descriptors. The training set is a subset of compounds that have had the property of interest measured in the laboratory. A regression fit is performed to relate the laboratory measurements of these compounds to the coefficients in the model. Validation and error assessment of the QSAR model is performed with the remaining laboratory measurements that were not included in the original training set. The model is generally valid for chemical compounds that are similar to those used in the training set.

The properties examined in the present study include 1) Boiling Point at 1 atm (101325 Pa), 2) Vapor Pressure at standard ambient temperature (25^o C), 3) solubility in water (mg/L), 4) –log(acid/base dissociation constant) in water (pK_a), and 5) the –log(octanol/water partitioning coefficient) (K_{ow}). For properties 1 to 3, EPI Suite version 4.11, and ACD/Labs/PhysChem 12.0 were used. For property 4 (pKa), ACD Labs, and ChemAxon's MARVIN pK_a and K_{ow} calculators were used to make predictions. For property 5, the K_{ow}, predictions were available from EPI Suite, ACD/Labs, MARVIN, and VEGA. These estimation models are based on group-contribution methods. An additional method called ADF COSMO-RS that is based on DFT methods was also used for comparison. Certain molecular fragments or functional groups tend to add to the magnitude of given properties. More frequent occurrences of these groups result in greater magnitudes for those properties, and some of the methods include correction factors that are summed into the property value based on the occurrence of certain atoms or other functional groups. For example, with pKa estimations, certain functional groups are known to be ionizable, such as amines. Occurances of these functional groups are known to be proportional to the experimental pKa, and as a result, these can be used to generate a knowledgeable guess.

Table 1: Molecular Descriptor Classes and Examples

Constitutional	Electronic/geometric	Physico-chemical	Fragment/Structure	Topological
# H-bonds	dipole moment	lipophilicity	functional groups	atomic branching
Hammett constants	molecular volume	polarizability		bonds to atom
# double bonds	bonds to atom			molecular shape index
molecular weight				polar surface area
# Rings				electrostatic field

2.1.1 Differences between EPI Suite/Mbbpvpwin and ACD Labs

The Mbbpvpwin component of EPI Suite utilizes the method of Stein & Brown[34] to estimate the boiling points. Using the standard QSAR approach, a linear model is used to estimate boiling point. Using the same 41 groups used by Joback[35] and Reid,[36] an additional 85 groups are added to the method. It should be noted that some of the additional groups are subgroups of the original 41. The training set consisted of 4426 different organic compounds, with an additional set of 6584 measurements for validation of the method.

In contrast to the method used in EPI Suite, the boiling point is not calculated directly using a QSAR approach, but rather the value of a function K. The ACD Labs User's Guide states that the boiling point follows a nonlinear form similar to the Antoine equation:

$$n_i = a_0 + \frac{a_1}{bp - a_2} \qquad \text{Equation (2)}$$

where n_i corresponds to the number of occurrences of group i, bp corresponds to the boiling point, a_0, a_1, and a_2 are empirically determined constants. However, they determined that the value of K, a function of the molecular volume (MV) and the boiling point (BP), is linearly dependent on the occurrences of given groups:

$$K = f(MV, BP) = c_0 + \sum_i c_{1,i}\, n_i \qquad \text{Equation (3)}$$

where n_i is the number of occurrences of a given group within a molecular structure, $c_{1,i}$ a weighting factor for that group that is determined by a regression of data, c_0 is a constant factor also determined from a regression fit of data. Because the algorithm is proprietary to ACD Labs, the form of the function relating the boiling point to K could not be found.

For vapor pressure, the most reliable method within EPI Suite is the modified Grain method shown in Lyman, which relates vapor pressure to the boiling point

as calculated above via a simple equation. This equation calculates vapor pressure from both solid and liquid materials.

2.1.2 Calculations Methods for the Octanol/Water Partition Coefficient

For this survey of predictive tools, a total of five software packages were available, including EPI Suite, ACD Labs, ChemAxon's Marvin, Vega, and ADF COSMO-RS. EPI Suite's KOWWIN is a QSAR based model with two regressions.[37-38] First, the molecule is divided into fragments based on core, non-hydrogen atoms. For 1120 different compounds with good experimentally determined K_{ow}, an initial regression yields weighting coefficients for the different fragments. A residue of errors remain, so correction factors are determined from more detailed grouping of the molecular fragments, taking into account structures such as rings or specific functional groups. The weighting for the correction factors are determined by a second regression with the full set of 2447 compounds with experimentally determined partition coefficients. The model within Vega is based on the same approach. The description of ACD Labs Log(P) algorithm in the User Guide is similar in that it is a QSAR based approach, but correction factors are not supplied. This algorithm assigns molecular fragments based on an internal database of 500 different functional groups as well as increments for different hybridizations of carbon atoms. Additional increments for 2000 intramolecular interactions such as ring structures and proximity to given functional groups are included. A relatively large training set of 18,412 chemical compounds is used. Chem Axon's log(P) (K_{ow}) calculator is QSAR based, utilizing a regression molecular fragment descriptors approach as described in Viswandhan.[39] This approach is augmented by including atomic partial charges, electron delocalization, ionic forms, and molecular polarizability. The model is also refined by additional molecular fragments. The model used in the Vega tool relies on the same approach by Viswandhan, and uses a training set of 2524 compounds.

2.1.3 Approaches for Calculating Water Solubilities

Two separate methods are available within EPI Suite. One method uses either an experimental or estimated K_{ow} to generate an estimate, and the second method uses the fragment based approach to get its estimate. In the first approach, an equation relating the water solubility to the K_{ow} and a fragment determined correction factor in Meylan & Howard[40] is used. If a reliable melting point temperature is available, then a second, similar equation is used that includes that quantity. The correction factors depend on the appearance of 15 different chemical functional groups, and a dataset of 1450 measurements of K_{ow} was used in the regression. The results of both approaches are shown here. ACD Labs references Meylan & Howard for its method as well, but no additional information could be located in the documentation.

2.1.4 Approaches for Calculating pK_a

As described in the software documentation, ACD Labs uses a fragment approach for calculating pK_a, and it relies on the presence of heteroatoms in the hydrocarbon structure for estimation. Hammett[41] first observed a simple equation to

describe the dependence of the pK_a of substituted compounds to the pK_a of the unsubstituted compound. ACD Labs uses parameterized Hammett-type equations to describe 1500 possible combinations of more than 650 ionizable functional groups. The change in pKa encountered when substituting different functional groups on the original ionizable fragment is encapsulated within the electronic substitutent constant, which is determined for over 1200 possible substituents. Additional corrections are added to account for the effect of distance through the hydrocarbon backbone between the ionizable center and a substituent, as well as through aliphatic and aromatic rings. ChemAxon's Marvin pK_a calculator uses a fragment based approach as well,[42-43] where pK_a depends linearly on fragment partial charge increments, polarizability increments, and structure specific increments, such as rings. Predictions are based on a regression relating these increments to the experimental data. Information on the data training set could not be located for the Marvin pK_a prediction tool.

2.2 Principle of Operation of COSMO-RS

The properties (boiling point, vapor pressure, K_{ow}, etc.) examined in this report are a reflection of the molecular solvation properties. For example, K_{ow} reflects the different energetics of solvation for a solute molecule in water and in octanol. Calculation of these energetics using an electronic structure method for both the solute molecule and sufficient solvent molecules to simulate solvation is still prohibitively computationally expensive. Fortunately, polarizable continuum models for solvents are proven to be reliable yet computationally tractable. As a result, it is possible to calculate solvation properties of a given molecule from the calculated chemical potential of the molecule in solution and the gas phase.. ADF COSMO-RS relies on the difference between the charge density on the molecular surface in vacuum and the charge density of the molecular surface within a polarizable continuum model.

Polarizable continuum models treat the solvent around a solute molecule as an infinite continuum with some dielectric constant. A cavity is carved out of this continuum to make way for the solute molecule. The solute molecule has some distribution of electrostatic charge contributing to a dipole moment. The polarizable continuum responds to this electrostatic charge resulting in an image charge that "screens" the charges in the solute molecule. While the solvent continuum will affect the electronic structure of the solvent molecule, the electronic structure can be refined until self consistency is obtained with the solvent continuum. There is an energy associated with this screening charge, and that screening charge serves as a descriptor for models of boiling point, vapor pressure, solubility, pK_a, miscibility, etc. Because the molecular descriptor is the result of a quantum mechanical calculation of a molecule, rather than from a group contribution method, we expect the approach to be more robust towards molecules containing groups not included in the original training set.[32-33]

For COSMO-RS, the cavity geometry is related to the radii of the atoms in the solute molecule. COSMO-RS is not a first principles method, because the atomic radii are fitted by a regression of 642 data points for a variety of properties, such as partition coefficients, and vapor pressure. The result is approximately 120% of the van der Waals

atomic radii. In the respect that certain degrees of freedom are fixed by a regression fit to experimental data, the COSMO-RS method some similarity to existing QSAR methods. A key difference is that instead of relying on the structure alone to make predictions, COSMO-RS calculates a molecular descriptor based on the charge distribution on the molecule within a polarizable continuum and the molecule in vacuum.

According to the ADF COSMO-RS Tutorial (http://www.scm.com/Doc/Doc2010/CRS/CRSGUI_tutorial/page48.html), pKa estimates can be made from four ADF calculations: 1) DFT gas-phase geometry optimization of the compound of interest, 2) COSMO-RS calculation of this optimized structure in an implicit water polarizable continuum solvent, 3) DFT gas-phase geometry optimization of the conjugate acid or base of the compound of interest, and 4) COSMO-RS calculation of the optimized conjugate structure in the continuum solvent. For multiprotic compounds, this four-step process can be repeated to individually calculate the pKa of each protonatable site, although the COSMO-RS parameterization was limited to monoprotic molecules only, so multiprotic results are expected to be inaccurate due to the large charges of the unparameterized ions.

At T = 298.15 K, ADF COSMO-RS uses the following equations when the compound of interest is an **acid** (deprotonation of acid -> conjugate base):

$$\boldsymbol{HA}\ (aq, 1M)\ +\ H_2O\ (l)\ \rightarrow\ H_3O^+\ (aq)\ +\ A^-\ (aq) \qquad \text{Equation (4)}$$

$$pKa = 0.62 * 0.733 * \Delta G * mol/kcal + 2.10 \qquad \text{Equation (5)}$$

$$\Delta G = G(conjugate\_base) - G(acid) + G(hydronium) - G(water) \qquad \text{Equation (6)}$$

ADF COSMO-RS uses the following equations when the compound is a **base** (deprotonation of conjugate acid -> base):

$$HB^+\ (aq, 1M)\ +\ H_2O\ (l)\ \rightarrow\ H_3O^+\ (aq)\ +\ \boldsymbol{B}\ (aq) \qquad \text{Equation (7)}$$

$$pKa = 0.67 * 0.733 * \Delta G * mol/kcal - 2.00 \qquad \text{Equation (8)}$$

$$\Delta G = G(base) - G(conjugate\_acid) + G(hydronium) - G(water) \qquad \text{Equation (9)}$$

Where,

G(hydronium) = -310.737 kcal/mol (ADF COSMO-RS calculation result)

G(water) = -332.353 kcal/mol (ADF COSMO-RS calculation result)

The ADF COSMO-RS calculations output the Gibbs free energy (G) values. Using the equations above and the G values obtained from ADF COSMO-RS, pKa values were calculated for each compound and conjugate pair. The acid set of equations gave the best (highest) pKa values, even when the compound of interest was a base. For example,

in the case of an amine species, the ammonium ion form (acid) was treated as the compound of interest, the amine form (base) was treated as the conjugate, and the acid parameter values above were used to calculate pKa. However, the reverse of this approach, ie, treating the amine as the compound and the ammonium as the conjugate and using the base parameter values, should have matched the ion charges better (HB^+ -> B rather than HA -> A^-), according to the parameterization. For glycerol, the middle OH group had a lower pK_a value and was assumed to deprotonate before either of the two end OH groups.

2.3 Overview of Traditional Chemical Warfare Agents and Simulants

Chemical warfare agents were initially used in the First World War. Compounds arising from that time period include sulfur mustard (HD) and the Lewisite compounds (L1, L2, and L3). Due to the similarities of its physical properties, glycerol is typically used as a simulant for mustard agents. Compounds representative of the Second World War time period include the nitrogen mustards, e.g. HN1, and organophosphate nerve agents tabun (GA), sarin (GB), soman (GD), and cyclosarin (GF). The V series agents (VG and VX) are representative of Cold War agents. Because of their similarity to organophosphate pesticides, common pesticides can serve as simulants for the G and V agents, such as DMMP, DIMP, metamidophos, and malathion. These compounds were considered to be somewhat representative of the chemical warfare agent threat and their chemical structures were used in the present study. Also, for most of the chemical properties of interest, e.g. solubility and K_{ow}, laboratory data was readily available.[11, 44-45] Because the fate of pharmaceuticals and dyes in the environment were also of interest, Disperse Red 9 and the cocaine molecule were considered.

3. METHODS

EPI Suite version 4.11 (KOWWIN v1.68, Mpbpwin v1.48, WaterNT v1.0, and WSKOWWIN v1.42), ACD Labs version 12.0, ChemAxon's pK_a and K_{ow} online calculators, and the VEGA Non-interactive Client for LogP (K_{ow}) predictions. To accommodate calculations of large numbers of compounds in a batch format, all predictive tools accept input files containing the structures in Simplified Molecular-Input Line Entry System (SMILES) notation, or the CAS numbers. Table 2 shows the list of traditional chemical warfare agents and simulants, as well as Disperse Red 9, a smoke grenade dye, along with their CAS numbers and structure in SMILES format that were used as input into the predictive tools. For ADF COSMO-RS, two quantum mechanical DFT calculations were necessary, one calculation simulating vacuum, the other simulating the molecule of interest embedded in a polarizable continuum (COSMO). The Becke '88 Perdew '86 (BP86) functional with the triple zeta TZP basis set were used for the quantum mechanical calculation.

Table 2: Compound Names, SMILES Structures, and CAS Numbers

Compound	SMILES structure	CAS #
glycerol	C(O)(CO)CO	56-81-5

HD (sulfur)	C(Cl)CsCCCl	505-60-2
HN1 (nitrogen mustard)	C(Cl)CN(CCCl)CC	000538-07-8
metamidophos	COP(N)(=O)SC	65960-97-6
malathion	C(=O)(C(CC(=O)OCC)SP(=S)(OC)OC)OCC	121-75-5
DMMP	COP(C)(=O)OC	756-79-6
DIMP	C(C)(C)OP(C)(=O)OC(C)C	169301-54-6
GA	C(#N)P(=O)(N(C)C)OCC	77-81-6
GB	CP(=O)(F)OC(C)C	107-44-8
GD	C(C)(C)(C)C(C)OP(C)(=O)F	96-64-0
GF	C1(OP(C)(=O)F)CCCCC1	329-99-7
L1	C(=CCl)[As](Cl)Cl	541-25-3
L2	C(=CCl)[As](Cl)C=CCl	40334-69-8
L3	C(=CCl)[As](C=CCl)C=CCl	40334-70-1
VX	C(C)(C)N(C(C)C)CCSP(C)(=O)OCC	50782-69-9
VG	C(CN(CC)CC)SP(=O)(OCC)OCC	78-53-5
Disperse Red 9	O=C2c1ccccc1C(=O)c3c2cccc3NC	82-38-2
cocaine	CN1[C@H]2CC[C@@H]1[C@H]([C@H](C2)OC(=O)C3=CC=CC=C3)C(=O)OC	50-36-2

4. RESULTS

4.1 Boiling Point

Table 3: Table of Predicted Boiling Points and Experimental[44-46] Measurements in degrees Celsius.

Compound	EPI	ACD	COSMO-RS	Experiment
glycerol	231	290	325	290
HD (sulfur)	210	216	265	216
HN1 (nitrogen mustard)	212	136	301	194
metamidophos	223	209	324	high*
malathion	351	385	701	high*
Dimethylmethylphosphonate	152	181	283	181
diisopropyl methyl phosphonate	210	214	394	high*
GA	267	240	324	240
GB	140	147	218	147

GD	183	201	306	198
GF	223	237	278	239
L1	156	203	215	196.6
L2	204	229	290	n/a
L3	247	215	318	n/a
VX	321	319	550	298
VG	337	315	613	vacuum*
disperse red-9	397	463	382	n/a
Cocaine	363	395	505	n/a
RMSE	29	20	107	

*Measurement performed under vacuum rather than 101 kPa or 760 mm/Hg, and as a result can't be compared to the prediction.
n/a. Experimental not available due to lack of measurement or decomposition or sublimation upon heating.

Table 3 compares predicted boiling points of the set of chemicals to available experimental measurements. For some experimental measurements, either the compounds of interest decomposed before reaching the boiling point, or the measurements were performed under vacuum and the boiling point was evaluated at low pressure (~0.2 mm/Hg). As a result, experimental measurements were unavailable for a number compounds, such as metamidophos, malathion, DIMP, and VG. For other compounds, such as Lewisite L2 and Lewisite L3, a measurement could not be located.

Using the available experimental measurements, it was possible to quantify the percent error of the predictions relative to the available experimental measurements. ACD/Labs method showed the lowest average RMSE error of 20^{o} C, compared to that of EPI Suite (29 RMSE error), and the ADF-COSMO-RS method (average error of 107^{o} C). For glycerol, HD (sulfur mustard), DMMP, GA, GB, GD, GF, and L1, the ACD Labs software almost exactly predicts the boiling point. ACD Labs overestimates the boiling point of VX by 7%. Given that ACD Labs takes into account the nonlinear dependence of the boiling point on number of occurrences of molecular fragments/groups, it does not seem surprising that it robustly predicts boiling points for the tradtional agents. Given the accuracy of ACD Labs predictions for the traditional agents, we conclude the traditional agents lie within the span of the models used by ACD Labs. EPI Suite also performs well, where the greatest error, underestimating the boiling point by 20%, is for Lewisite L1. This is to be expected given that the Lewisites contain the semi-metal Arsenic, which is outside the training set of EPI Suite.

The predictions obtained from ADF COSMO-RS deviated from the experimental measurements the most. For compounds containing a tertiary amine such as VX or nitrogen mustard (HN1), the predictions overestimated by more than 100^{o} C. Predictions for the organophosphates also differed from experiment by more than 100^{o} C, except for GA and GF, which differed from experiment by 80^{o} C and 40^{o} C, respectively. Although ADF COSMO-RS is not a traditional QSAR in the sense that a regression is performed to relate the occurrences of some descriptor to the property of interest, there is still an empirical fit of the atomic radii used to define the molecular cavity within the polarizable continuum. It is quite possible the training set used in the regression is

insufficient and the traditional agents are outside of the domain of applicability. On the other hand, it is possible the degrees of freedom afforded by the atomic radii fit do not result in a sufficiently expansive applicability domain. Lastly, although Disperse Red 9 and cocaine have secondary amine functional groups and can exist in a salt form, neither EPI Suite nor ACD Labs can make predictions for the ionic form.

4.2 Vapor Pressure

Table 4: Vapor Pressures in Pa at 25^{o} C

	EPI	ACD	COSMO-RS	Experiment
glycerol	1.1E-02	<1.0E-01	1.0E-03	2.2E-02
HD (sulfur)	2.1E+01	2.7E+01	3.0E+01	1.5E+01
HN1 (nitrogen mustard)	2.6E+01	1.1E+03	1.3E+01	3.3E+01
metamidophos	9.1E+00	2.7E+01	2.6E-02	4.7E-03
malathion	1.7E-02	<1.0E-01	0.0E+00	4.5E-04
Dimethylmethylphosphonate	1.2E+02	1.3E+02	1.3E+01	1.3E+02
diisopropyl methyl phosphonate	3.0E+01	2.7E+01	4.6E-01	3.7E+00
GA	6.2E+00	5.3E+00	3.3E+00	9.3E+00
GB	6.1E+02	8.0E+02	1.4E+02	3.8E+02
GD	5.3E+01	6.7E+01	9.2E+00	5.3E+01
GF	6.5E+00	9.3E+00	2.3E+01	5.9E+00
L1	7.9E+01	5.3E+01	2.0E+02	7.7E+01
L2	3.9E+01	1.3E+01	1.6E+01	
L3	3.6E+00	2.7E+01	8.8E+00	
VX	2.9E-01	<1.0E-01	4.0E-03	9.3E-02
VG	3.6E-02	<1.0E-01	5.0E-03	3.5E-02
disperse Red-9	4.1E-05	<1.0E-01	7.1E-01	9.3E-07
cocaine	1.7E-03	<1.0E-01	1.2E-02	3.9E-05

The values within Table 4 compare vapor pressure predictions to experimental values. Unlike the predictions for boiling point, the deviations from measurements vary from factors of 2 to many orders of magnitude. For most of the predictions, the large magnitude of error is not significant. This is especially true when both the prediction and the measurement are very small, as in the comparison between prediction and experiment for glycerol, malathion, VG, and Disperse Red 9. Many of the predictions for all three tools are reasonably close to the experimental measurement, such as for HD (all within a factor of 2 or less), HN1, GA, GB, and L1. On the other hand, some of the deviations between prediction and measurement are significant, and could

drastically affect the output of fate models using the predicted vapor pressures. The most significant deviation on the order of two to four orders of magnitude occured for both EPI Suite and ACD Labs with metamidophos. For EPI Suite, the $-NH_2$ fragment in metamidophos is treated the same as a typical amine group, and the P-N linkage is not recognized. Most likely, the amine group in metamidophos is much different than a typical amine. For GA, although there is an N-P linkage, it is possible the effect of that linkage is limited by the fact that the nitrogen atom is joined to two ethyl groups. A fate model may predict rapid dissipation for this compound using the predicted vapor pressures, when in fact they are persistent. There is also a limit to the predictive range of ACD Labs. A vapor pressure lower than 0.1 Pa is outside of the predictive range of ACD Labs. Thus, ACD Labs can provide a lower bound for vapor pressure, but only for vapor pressures higher than 0.1 Pa.

ADF COSMO-RS performed reasonable well against all compounds in the set. Except for DMMP and DIMP, the predictions were well within an order of magnitude of the measurements. Measurements for low vapor pressure compounds such as VX, VG, and disperse Red 9, were orders of magnitude different from the predictions, but for these compounds, both the prediction and measurement were very small, impacting a fate model minimally. Given that for several organophosphorous compounds, ADF COSMO-RS predictions matched experiment well (G and V agents, and the pesticides malathion and metamidophos), it is surprising that this method should encounter difficulty with DIMP and DMMP where the prediction varies by more than a factor of ten with respect to the experimental measurement. It should be noted that DMMP and DIMP are very symmetric molecules. As a result, we expect polarity and polarizability to depend strongly on the molecular conformation. For ADF COSMO-RS, properties are calculated at the lowest energy conformation, where at typical laboratory ambient temperatures, the molecules may be sampling many more conformations. As a result, predictions of properties that depend on polarity and polarizability may differ greatly from laboratory measurements. At the same time, ADF COSMO-RS does not seem to deviate much more than an order of magnitude in its vapor pressure estimate, where for a number of instances, both EPI Suite and ACD Labs deviate for up to 4 orders of magnitude in vapor pressure. Deviations of that magnitude can cause serious errors in predictions of the fate and transport of these materials in the environment.

4.3 Log Water/Octanol Partition coefficient (K_{ow})

Table 5: Comparison of Kow (-log(P)) predictions from EPI Suite, ACD Labs, Marvin Kow Calculator, Vega, COSMO-RS, and experimental values for the selected traditional chemical warfare agents and simulants.

Compound	EPI	ACD	MARVIN	Vega	COSMO	EXP
glycerol	-1.7	-1.9	-1.8	-2.3	-1.7	-1.76
HD (sulfur)	2.4	2.1	2.0	3.2	2.8	1.37
HN1 (nitrogen mustard)	1.4	1.4	1.9	2.0	3.6	2.02
metamidophos	-0.9	-0.8	-0.3	-0.9	0.1	-0.8
malathion	2.3	2.4	1.5	3.0	3.6	2.36

Dimethylmethylphosphonate	-0.6	-0.9	-0.1	-0.6	-0.5	-0.61
diisopropyl methyl phosphonate	1.2	0.8	1.4	1.2	1.4	1.03
GA	0.3	0.1	-0.1	-0.2	1.6	0.38
GB	0.3	0.5	0.8	-0.2	1.1	0.3
GD	1.7	1.8	2.1	1.2	2.3	1.78
GF	1.6	1.6	1.7	1.2	2.2	na
L1	2.6	n/a	2.5	1.9	2.7	2.56
L2	3.5	n/a	3.5	3.2	3.5	
L3	4.5	n/a	3.7	4.5	4.5	
VX	2.1	2.1	2.0	2.1	4.2	2.09
VG	1.7	2.9	1.8	1.7	4.8	1.7
disperse red-9	4.1	3.0	3.0	3.0	3.6	4.1
cocaine	2.2	2.3	2.3	2.5	3.9	2.3
RMSE	0.3	0.5	0.5	0.7	1.3	

Table 5 shows a comparison of the predictions of octanol/water partition coefficients from various available QSAR tools compared to reported laboratory measurements in –log units. Laboratory measurements could not be found by the authors for cyclosarin (GF) or two of the Lewisite agents L2 and L3. When taking the square root of the average of the squares of differences between the predictions and the experimental measurements in –log units, one is able to perform a rough ranking of the different tools. EPI Suite had the lowest average error (in –log units) of 0.33. ChemAxon's Marvin –log(P) Calculator ranked next with an average error of 0.50. The average error for the predictions from ACD Labs was close with a value of 0.54. The average error for the Vega –log(P) Java based calculator was 0.7. The largest deviation from experimental value is obtained between the predictions from ADF COSMO-RS for this particular set of compounds.

For EPI Suite, ACD Labs, Marvin, and Vega, sulfur mustard (HD) yields the largest error, followed by the prediction for nitrogen mustard (HN1) for EPI Suite. Examining the EPI Suite output, it appears only basic atomic fragments of Cl, thio-ether, and methylene carbon groups are recognized. The proximity of the chlorine atoms to the sulfur are not accounted for in the available model. The error for HN1 is a factor of 3 less, and an additional correction factor for the ClCCNCCCl fragment appears. Clearly, the correction factors arising from the second regression of data and deviations from data from the initial QSAR prediction in EPI Suite's KOWWIN tool is quite effective. Another point to make is the molecular symmetry of HD and HN1. The symmetry is not recognized in the fragment output of EPI Suite, or any of the other fragment based methods, where the fragment values are blindly summed into the property prediction. K_{ow} should be affected by molecular polarity, and the contribution of polar groups to the molecular polarity can be cancelled out by symmetry. As a result, a failure to account for molecular symmetry may result in inaccuracy.

ChemAxon's Marvin Log(P) Calculator also uses additional correction factors derived from molecular descriptors such as atomic partial charges and polarizability. HD and HN1 also exhibit error for ACD Labs, but this error is less than the predictions from EPI Suite (0.593 and 0.36, respectively). The reduced error for these two compounds could be attributed to the large training set used for the ACD Labs algorithm. However, for VG and Disperse Red 9, errors greater than a log unit are obtained, so that for the limited set of compounds considered, ACD Labs exhibits slightly more error than EPI Suite and ACD Labs. Although Vega is based on the same approach as that used in EPI Suite's KOWWIN program, the average error exhibited is greater than the predictions from EPI Suite. An examination of the greatest errors show HD gives the largest magnitude error on the order of nearly two log units, and Disperse Red 9 an error of 1 log unit. The output of Vega does indicate that HD and Disperse Red 9 are well outside of the applicability range. Interestingly, the prediction from Vega is identical to the output from EPI Suite with the correction factor of 1.1000 subtracted. Most likely the difference between Vega and EPI Suite is the implementation of the correction factors. The greatest error is consistently obtained with the predictions of ADF COSMO-RS on the limited data set studied here. We conjecture the error can be attributed to the fact that empirical regression of the ADF COSMO-RS method is performed on the COSMO atomic radii rather than quantum mechanically calculated molecular descriptors, such as charge distribution, HOMO-LUMO difference, etc.

4.4 Water Solubility

The predictions from EPI Suite, ACD Labs, and ADF/COSMO are compared to experimental values in Table 6. To estimate error, the root mean squares of the difference between the log of the predicted solubility and the log of the experimental solubility were used to rank the different methods. Both of the methods within EPI Suite (K_{ow} and fragment based) and ACD Labs had nearly equivalent errors of 0.87, 1.1, and 1.0, respectively, although the method within EPI Suite that relied on K_{ow} appeared to have the least error. ADF COSMO-RS had a slightly larger error with an average of the difference between the logarithm of the prediction and the logarithm of the measurement of 1.3.

For EPI Suite's K_{ow} method, the four compounds that result in the greatest deviation from experiment is HN1, sarin (GB), soman (GD), and VX. For the greatest magnitude error, with HN1, solubility was overestimated by nearly six orders of magnitude. For triethylamine, an analog of HN1 without the two chlorine atoms, EPI Suite's prediction of 6.82×10^4 mg/L is very close to the experimental measurement of 6.86×10^4 mg/L.[47] Clearly, the chlorine atoms in HN1 have a drastic effect on solubility that is not accounted for in EPI Suite's model. This is surprising since the prediction of K_{ow} from EPI Suite is less than 0.5 log units different than experiment. One thing to be noted in the EPI Suite output is that only one correction factor is applied due to the aliphatic amine group. No correction factors for chlorine are included. Another issue to mention is that molecular symmetry can play a role. Although chlorine atoms within hydrocarbon molecules tend to be locally polar, if the arrangement of chlorine atoms is symmetric, that polarity may cancel out resulting in a nonpolar molecule overall. In contrast to HN1, the GB, GD, and VX solubilities are underestimated by about an order of

magnitude. Within the output, it appears no correction factors are available for the unique structures around the phosphorus in the G and V agents, and so the closeness between measurement and prediction for GA may be fortuitous.

For the fragment based method in EPI Suite, the method seems to consistently overestimate solubility in water. DIMP, GF, Disperse Red 9, and HN1 yield the greatest differences between experiment and prediction. As with the K_{ow} method in EPI Suite, the structure around phosphorus atom seems to be unique for GF. Although all portions of DIMP appear to be represented in the chosen fragments for property estimation, the predicted solubility for DIMP is still orders of magnitude greater than the experiment. It is possible that the discrepancy arises from the fact that the methyl group attached to the phosphorus is counted in the same way as the methyl and methylene groups in isopropyl fragments. However, there is a similar methyl group in GB, and both experiment and prediction show GB highly soluble in water. Molecular symmetry may be an explanation for less than expected water solubility, where symmetry cancels out molecular polarity. For GF, the large difference between prediction and experiment is not clear either. All portions in the GF molecule also have corresponding fragments in the model, yet there is poor agreement. At first guess, the unsaturated ring would reduce the solubility, but those aliphatic carbons are accounted for within the model. An evaluation of the solubility of cyclohexanol shows good agreement between experiment and prediction (4.2×10^4 vs. 4.7×10^4 mg/L, respectively). Again, the substitutents around the phosphorus atom are not unique, and molecular symmetry is not an issue. Perhaps there is some interaction between the aliphatic carbons and the substituents around the phosphorus atom. The predictions for Disperse Red 9 and HN1 also overestimate the solubility by about an order of magnitude. Both of these molecules have few polar or ionizable functional groups. Molecular symmetry that cancels the effect of polar fragments may reduce the overall polarity of the molecules in water, leading to errors in the prediction.

For the ACD Labs predictions, Disperse Red 9, HN1, DIMP, and GF also had the greatest differences between prediction and measurement. However, unlike EPI Suite, the version of ACD Labs we possess did not report the details of the fragment contribution. Since ACD Labs uses the same approach as the EPI Suite Fragment method, we can assume some of the same issues affect this software package.

Table 6: Water Solubility (mg/L)

Compound	Solubility (mg/L)				
	EPI Kow	EPI Frag	ACD	COSMO-RS	Experiment
glycerol	1.0E+06	1.0E+06	7.1E+05	1.0E+06	1.0E+06
HD (sulfur)	6.1E+02	4.3E+02	4.2E+03	3.7E+02	6.8E+02
HN1 (nitrogen mustard)	4.0E+04	7.3E+03	1.6E+04	1.3E+02	1.6E+02
metamidophos	4.0E+05	1.0E+06	1.0E+06	1.0E+06	1.0E+06
malathion	7.8E+01	4.3E+02	3.0E+02	6.6E+01	1.4E+02
DMMP	3.2E+05	1.0E+06	6.3E+05	1.0E+06	1.0E+06
DIMP	7.3E+03	2.2E+05	3.4E+04	1.0E+06	1.5E+03

GA	3.2E+04	1.0E+06	3.5E+05	1.6E+04	9.8E+04
GB	4.6E+04	1.0E+06	4.2E+05	1.0E+06	1.0E+06
GD	1.6E+03	3.4E+05	5.0E+04	1.1E+04	2.1E+04
GF	2.1E+03	5.4E+05	6.2E+04	1.3E+04	3.7E+03
L1	2.6E+02	4.7E+03	n/a	1.8E+03	5.0E+02
L2	2.9E+01	5.0E+02	n/a	2.0E+02	
L3	3.3E+00	5.2E+01	n/a	1.9E+01	
VX	3.2E+03	9.1E+04	4.8E+03	5.0E+02	3.0E+04
VG	6.6E+03	4.7E+04	1.6E+04	1.0E+02	3.0E+04
disperse red-9	6.8E-01	8.6E+00	4.5E-04	1.7E+02	1.2E-01
cocaine	1.3E+03	1.0E+03	5.3E+02	3.5E+02	1.8E+03
RMSE*	0.87	1.1	1.0	1.3	

*Root Mean Square of the differences between log(predicted) and log(experiment).

4.5 pK_a Predictions

Table 7 contains a comparison of pKa predictions from ACD Labs, ADF/COSMO-RS and Marvin. Unfortunately, for many of the compounds, we could not locate most of the pKa values for the compounds except for glycerol, HN1 nitrogen mustard, and VX. For glycerol, only the first proton dissociation constant was reported at 14.15 log units. There was less than 0.5 log units difference between ACD Labs, Marvin, and the experimental measurement. ADF/COSMO was two log units off. The same was true for HN1 with both Marvin and ACD Labs less than 0.2 log units off, and ADF/COSMO-RS was 0.83 log units off. For VX, the errors were greater, with ADF/COSMO-RS 0.7 log units off, ACD Labs 1.2 log units off, and Marvin 2.0 log units off. Good agreement was determined for glycerol, where the R-OH functional group is contained within the ACD Labs library of ionizable functional groups. A similar degree of accuracy is found with Marvin, although a slightly different approach is used. ACD Labs applies an empirical correction factor, and perhaps the contribution from partial atomic charge and polarizability in Marvin acts in the same way as the correction factor within ACD Labs. Although the predictions from ADF/COSMO-RS were somewhat close to the experimental values, the difference between the predictions from this method and experimental results were consistently greater than the results from ACD Labs or Marvin. It should also be noted that Marvin was unable to make a prediction for GA and metamidophos. Most likely it could not recognize the N-P linkage in either GA or metamidophos.

Table 7: Comparison of pKa predictions and available experimental data.

	ACD			ADF COSMO-RS			Marvin		EXP	
	H1	H2	H3	H1	H2	H3				
glycerol	13.7	14.8	15.9	12.3	14.4	21.9	13.6	15.2	14.15	

HD (sulfur)	n/a						n/a			
HN1 (nitrogen mustard)	6.5			7.4			6.3		6.57	
metamidophos	-0.6			3.2			na			
malathion	n/a						n/a			
DMMP	n/a						n/a			
diisopropyl methyl phosphonate	n/a						n/a			
GA	-4.7			-2.4			n/a			
GB	n/a						n/a			
GD	n/a						n/a			
GF	n/a						n/a			
L1	n/a						n/a			
L2							n/a			
L3							n/a			
VX	9.8			7.9			10.6		8.6	9.4
VG	9.4			6.2			9.9			
disperse red-9	2.3			2.3			2.2			
cocaine	9.0			9.9			8.9			

For most of the compounds, such as HD, malathion, DMMP, DIMP, GB, GD, GF, L1, L2, and L3, a readily ionizable atom center was not found within the molecule, and a pKa could not be determined for these compounds. The rest of the molecules possessed either an amine functional group, or an oxygen atom with a proton. Although ADF/COSMO-RS is not based on the fragment QSAR method, this approach did not determine a pKa.

5. CONCLUSIONS.

A number of in-silico tools were used to predict physico-chemical properties that are relevant to modeling of environmental fate of a number of traditional agents and simulants, and these results were compared to available experimental data. Specifically, the boiling point, vapor pressure, log of the water/octanol partitioning coefficient (K_{ow}), water solubility, and pK_a were the properties surveyed. For boiling point, both ACD Labs and EPI Suite were highly accurate to within 20^o C and 29^o C. EPI Suite, ACD Labs, and ADF COSMO-RS performed quite well for vapor pressure predictions, except that ACD Labs could not generate predictions for vapor pressures of less than 0.1 Pa. For K_{ow}, EPI Suite, ACD Labs, ChemAxon's Marvin, Vega, and ADF COSMO-RS were evaluated. When considering the RMSE for the K_{ow} values, EPI Suite, ACD Labs, Marvin, Vega, and ADF COSMO-RS resulted in 0.3, 0.5, 0.5, 0.7, and 1.3 log units, respectively. ACD Labs was unable to give a prediction for any of the Lewisite compounds. EPI Suite's K_{ow} estimation method proved to have the smallest difference between experiment and measurement. When considering the root mean square

differences between the logs of the predictions and measurements, EPI Suite's K_{ow} method, EPI Suite's Fragment based estimation method, ACD Labs, and ADF COSMO-RS had values of 0.87, 1.1, 1.0, and 1.4, respectively. We hypothesize that the greatest difference between prediction and experimental measurement occurred for ADF COSMO-RS because although it is based on DFT calculations, there is still an empirical fit that is based on only 642 compounds. There is not enough information to comment on the applicability domain for ADF COSMO-RS. Also, ADF COSMO-RS depends on a single molecular descriptor, i.e. the difference in charge density between gas phase and condensed phase calculations. We would expect a regression on a larger training set including additional molecular descriptors to greatly improve performance of this approach. For pKa estimations, experimental data for only three of the compounds examined could be located, and so the accuracy of ACD Labs and ChemAxon's Marvin towards the traditional agents could not be evaluated.

We reason that two issues contribute to the difference between experimental measurements and estimations from fragment based methods. First, as would be expected, compounds containing elements or functional groups outside of the method's training set contributed to the average error, such as the element arsenic in L1. The same issue is true for vapor pressure, K_{ow}, and solubility calculations, where unique functional groups around the phosphorus atoms, such as the P-N chemical link to a primary amine in metamidophos results in error. For a property such as boiling point or vapor pressure, where constitutional descriptors such as the elemental constituents or molecular mass, details of the molecular structure are not important. For properties involving behavior in water, where molecular interactions are important, descriptors such as polarity or polarizability intuitively also become important. Yet, the fragment based methods treat each fragment the same regardless if it is attached to an aliphatic carbon atom or a phosphorus atom.

The second issue that seems apparent in examination of the data is that molecules that tend to produce the largest differences between model prediction and experimental measurement have molecular symmetry. For example, HN1 is highly symmetric around the nitrogen atom, and the carbonyl groups in Disperse Red 9 are arranged on opposite sides of a ring. For properties highly dependent on molecular structure and polarity, such as water solubility, a fragment based method can contribute significantly to the error, since the fragment contributions are treated additively. It is possible symmetry may cancel out the contributions from a given fragment, so that a property is overestimated, as occurs for EPI Suite's fragment method for solubility in water on HN1, DIMP, and Disperse Red 9.

Overall, EPI Suite and ACD Labs gave reasonable property predictions when the compound of interest was contained within the model applicability domains and were asymmetric. We project that a model based on quantum mechanical descriptors would be insensitive to these effects since 1) any unique fragment or functional group could be "translated" into a more universal property descriptor such as dipole moment, and 2) symmetry would be automatically accounted for.

LITERATURE CITED

(1) Bennett, E. R.; Clausen, J.; Linkov, E.; Linkov, I., Predicting physical properties of emerging compounds with limited physical and chemical data: QSAR model uncertainty and applicability to military munitions, *Chemosphere* **2009,** *77* (10), 1412-1418.

(2) Benfenati, E., The CAESAR Project for In-silico Models for the REACH Legislation, *Chemistry Central Journal* **2010,** *4(Suppl 1)*, I1.

(3) Opresko, D. M.; Hauschild, V. *Derivation of Health-Based Environmental Screening Levels for Chemical Warfare Agents*; U. S. Army Center for Health Promotion and Preventative Medicine: Aberdeen Proving Ground, MD, 1999.

(4) Bartelt-Hunt, S. L.; Barlaz, M. A.; Knappe, D. R. U.; Kjeldsen, P., Fate of Chemical Warfare Agents and Toxic Industrial Chemicals in Landfills, *Environmental Science & Technology* **2006,** *40*, 4219-4225.

(5) Czerwinski, S. E.; Skvorak, J. P.; Maxwell, D. M.; Lenz, D. E.; Baskin, S. I., Effect of Octanol:Water Partition Coefficients of Organophosphorus Compounds on Biodistribution and Percutaneous Toxicity, *Journal of Biochemical and Molecular Toxicology* **2006,** *20* (5), 241-246.

(6) Khalil, M. A. K.; Rasmussen, R. A. In *Modeling chemical transport and mass balances in the atmosphere*, CRC: 1985; pp 21-54.

(7) Neely, W. B.; Blau, G. E. In *Introduction to environmental exposure from chemicals*, CRC: 1985; pp 1-11.

(8) Karickhoff, S. W. In *Pollutant sorption in environmental systems*, CRC: 1985; pp 49-64.

(9) Schnoor, J. L. In *Modeling chemical transport in lakes, rivers, and estuarine systems*, CRC: 1985; pp 55-73.

(10) Britton, K. B. *Low temperature effects on sorption, hydrolysis and photolysis of organophosphonates: a literature review*; Univ. New Hampshire: 1986; p 82 pp.

(11) Marrs, T. T., *Chemical Warfare Agents: Toxicology and Treatment*. John Wiley & Sons: Chippenham, England, 2007.

(12) Benschop, H. P.; De Jong, L. P. A. In *Toxicokinetics of soman in rats, guinea pigs, and marmosets*, Swed. Def. Res. Establ.: 1989; pp 163-71.

(13) Lyman, W. J. In *Estimation of physical properties*, CRC: 1985; pp 13-47.

(14) Scheunert, I.; Vockel, D.; Schmitzer, J.; Viswanathan, R.; Klein, W.; Korte, F., FATE OF CHEMICALS IN PLANT-SOIL SYSTEMS - COMPARISON OF LABORATORY TEST DATA WITH RESULTS OF OPEN AIR LONG-TERM EXPERIMENTS, *Ecotoxicology and Environmental Safety* **1983,** *7* (4), 390-399.

(15) Xiao, F.; Gulliver, J. S.; Simcik, M. F., Predicting aqueous solubility of environmentally relevant compounds from molecular features: A simple but highly effective four-dimensional model based on Project to Latent Structures, *Water Research* **2013,** *47* (14), 5362-5370.

(16) Olkowska, E.; Ruman, M.; Polkowska, Z., Occurrence of Surface Active Agents in the Environment, *Journal of Analytical Methods in Chemistry* **2014**.

(17) Bintein, S.; Devillers, J., QSAR FOR ORGANIC-CHEMICAL SORPTION IN SOILS AND SEDIMENTS, *Chemosphere* **1994,** *28* (6), 1171-1188.

(18) Oman, C., Comparison between the predicted fate of organic compounds in landfills and the actual emissions, *Environmental Science & Technology* **2001,** *35* (1), 232-239.

(19) Jones, O. A. H.; Voulvoulis, N.; Lester, J. N., Aquatic environmental assessment of the top 25 English prescription pharmaceuticals, *Water Research* **2002,** *36* (20), 5013-5022.

(20) Wang, C.; Song, C.; Li, H.; Li, Z., *Effect of pH and cation strength on lincomycin sorption from water by soils*. 2009; p 278-286.

(21) Halfon, E.; Galassi, S.; Bruggemann, R.; Provini, A., Selection of priority properties to assess environmental hazard of pesticides, *Chemosphere* **1996,** *33* (8), 1543-1562.

(22) Eisenberg, J. N. S.; Bennett, D. H.; McKone, T. E., Chemical dynamics of persistent organic pollutants: A sensitivity analysis relating soil concentration levels to atmospheric emissions, *Environmental Science & Technology* **1998,** *32* (1), 115-123.

(23) Hippelein, M.; McLachlan, M. S., Soil/air partitioning of semivolatile organic compounds. 1. Method development and influence of physical-chemical properties, *Environmental Science & Technology* **1998,** *32* (2), 310-316.

(24) Bouraoui, F.; Boesten, J.; Jarvis, N.; Bidoglio, G., *Testing the PEARL model in the Netherlands and Sweden*. 2003; p 527-534.

(25) Simunek, J.; Jacques, D.; Langergraber, G.; Bradford, S. A.; Sejna, M.; van Genuchten, M. T., Numerical Modeling of Contaminant Transport Using HYDRUS and its Specialized Modules, *Journal of the Indian Institute of Science* **2013,** *93* (2), 265-284.

(26) Ruark, C. D.; Hack, C. E.; Robinson, P. J.; Anderson, P. E.; Gearhart, J. M., Quantitative structure-activity relationships for organophosphates binding to acetylcholinesterase, *Archives of Toxicology* **2013,** *87* (2), 281-289.

(27) Debruijn, J.; Hermens, J., QUALITATIVE AND QUANTITATIVE MODELING OF TOXIC EFFECTS OF ORGANOPHOSPHOROUS COMPOUNDS TO FISH, *Science of the Total Environment* **1991,** *109*, 441-455.

(28) Veith, G. D.; Mekenyan, O. G., A QSAR APPROACH FOR ESTIMATING THE AQUATIC TOXICITY OF SOFT ELECTROPHILES QSAR FOR SOFT ELECTROPHILES, *Quantitative Structure-Activity Relationships* **1993,** *12* (4), 349-356.

(29) Lill, M. A.; Dobler, M.; Vedani, A., In-silico prediction of receptor-mediated environmental toxic phenomena - Application to endocrine disruption, *Sar and Qsar in Environmental Research* **2005,** *16* (1-2), 149-169.

(30) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O., A stepwise approach for defining the applicability domain of SAR and QSAR models, *Journal of Chemical Information and Modeling* **2005,** *45* (4), 839-849.

(31) Lill, M. A., Multi-dimensional QSAR in drug discovery, *Drug Discovery Today* **2007,** *12* (23-24), 1013-1017.

(32) Klamt, A.; Jonas, V.; Buerger, T.; Lohrenz, J. C., Refinement and Parameterization of COSMO-RS, *Journal of Physical Chemistry A* **1998,** *102* (26), 5074-5085.

(33) Klamt, A., *COSMO-RS From Quantum Chemistry to Fluid Phase Thermodynamics and Drug Design*. Elsevier: Amsterdam, 2005.

(34) Stein, S. E.; Brown, R. L., Estimation of normal boiling point from group contributions, *Journal of Chemical Information and Computer Science* **1994,** *34*, 581-587.

(35) Joback, K. G. A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques. Massachusetts Institute of Technology, 1982.

(36) Reid, R. C.; Prausnitz, J. M.; Poling, B. E., *The Properties of Gases and Liquids*. Fourth ed.; McGraw-Hill, Inc.: New York, 1987.

(37) Hansch, C.; Leo, A.; Hoekman, D., *Exploring QSAR. Hydrophobic, Electronic, and Steric Constants. ACS Professional Reference Book*. American Chemical Society: Washington, DC, 1995.

(38) Meylan, W. M.; Howard, P. H., Atom/fragment Contribution Method for Estimating Octanol-Water Partition Coefficients, *Journal of Pharmaceutical Sciences* **1995,** *84* (1), 83-92.

(39) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K., Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameteres for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics, *Journal of Chemical Information and Computer Science* **1989,** *29*, 163-172.

(40) Meylan, W. M.; Howard, P. H.; Boethling, R. S., Improved method for estimating water solubility from octanol/water partition coefficient, *Environ. Toxicol. Chem.* **1996,** *15* (Copyright (C) 2014 American Chemical Society (ACS). All Rights Reserved.), 100-6.

(41) Hammett, L. P., The effect of structure upon the reactions of organic compounds benzene derivatives, *Journal of the American Chemical Society* **1937,** *59*, 96-103.

(42) Szegezdi, J.; Csizmadia, F., Prediction of dissociation constant using microconstants, *Abstracts Of Papers Of The American Chemical Society* **2004,** *227*, U1019-U1019.

(43) Szegezdi, J.; Csizmadia, F., Method for calculating the pKa values of small and large molecules, *Abstracts Of Papers Of The American Chemical Society* **2007,** *233*.

(44) Gupta, R. C., *Handbook of Toxicology of Chemical Warfare Agents*. Elsevier: New York, NY, 2009.

(45) Pohanish, R. P., *Sittig's Handbook of Toxic and Hazardous Chemicals and Carcinogens*. Elsevier: New York, NY, 2012.

(46) *CRC Handbook of Chemistry and Physics*. 71st ed.; CRC Press: Boston, 1991.

(47) Christie, A. O.; Crisp, D. J., ACTIVITY COEFFICIENTS OF N-PRIMARY SECONDARY AND TERTIARY ALIPHATIC AMINES IN AQUEOUS SOLUTION, *Journal of Applied Chemistry* **1967,** *17* (1), 11-+.

ACRONYMS

ACD	Advanced Chemistry Development (Labs)
ADF	Amsterdam Density Functional Code
CAS	Chemical Abstracts Service
CWA	Chemical Warfare Agents
COSMO	COnductor-like Screening Model
DIMP	diisopropyl methyl phosphonate
DMMP	dimethyl methyl phosphonate
DFT	Density Functional Theory
EPA	Environmental Protection Agency
GA	Tabun
GB	Sarin
GD	Soman
GF	Cyclosarin
HD	Sulfur mustard agent
HN1	Nitrogen mustard agent $(ClC_2H_4)_2N(C_2H_5)$)
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
L1	Lewisite ($AsC_2H_2Cl_3$)
L2	Lewisite ($As(C_2H_2Cl)_2Cl$
L3	Lewisite ($As(C_2H_2Cl)_3$)
QSAR	Quantitative Structure Activity Relationship
RMSE	Root Mean Square Error
SMILES	Simplified Moleclar-Input Line-Entry System
TIC	Toxic Industrial Chemical
VG	Amiton/Tetram, ($C_{10}H_{24}O_3PS$)
VX	Cold War Agent, ($C_{11}H_{26}NO_2PS$)


EDGEWOOD CHEMICAL BIOLOGICAL CENTER
A U.S. Army RDECOM Laboratory